REGRESSION

Zack Kertcher

Statistics for Management
Fall 2016

Plan for today

1. Quiz
2. Regression
3. Logistics

Regression

Regression is typically used for the following purposes:

- ✔ Identify relationships among IVs and DV
- ✔ Identify the form and magnitude of these relationships
- ✔ Predict the DV using IVs
- ✔ Be able to assess the accuracy of the model

Regression varieties

Regression Type	Use
Simple linear	Predicting a numeric DV using a single numeric IV
Multiple linear	Predicting a numeric DV using a multiple IVs, numeric and/or factors
Logistic	Predicting a binary factor DV using one/multiple IVs
Multilevel	Predicting one/multiple DVs using one/multiple IVs
Poisson	Predicting a DV for count data using one/multiple IVs
Time series	Modelling time series data
Polynomial	Predicting a numeric DV using a numeric IV, when the relationship is modeled as an *n*th degree polynomial.
Nonparametric	Predicting a numeric DV when model is derived from the data and not specified a priori
Robust	Predicting a numeric DV with one/multiple IVs, using a model that is resistant to outliers
.....	**Various other uses. R has > 200 regression functions!**

Simple (and multiple) linear regression

$$\hat{y} = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_n X_{\mathrm{ki}}$$

Predicted

Intercept When all IV values = 0

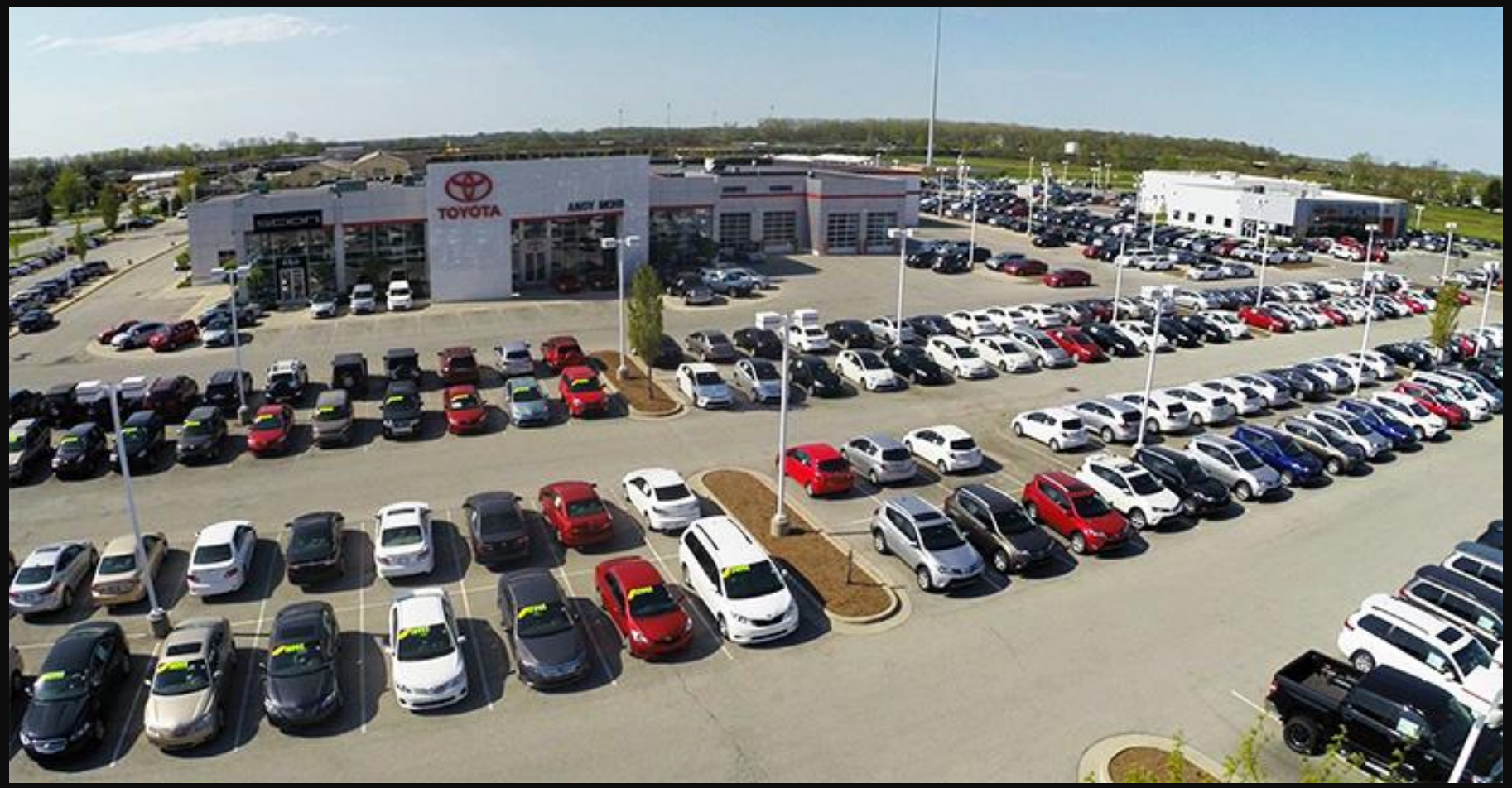
Coefficient* value of X_k (*i*th observation)

Linear regression assumptions

- ✔ The DV is normally distributed
- ✔ Y (IV) values are independent
- ✔ There is a linear relationship between the DV and IVs
- ✔ There is constant variance, such that the variance of the DV does not change with the levels of IVs (homoscedasticity)

And… we also check for multicollinearity (the extent to which two or more IVs in the model are highly correlated).

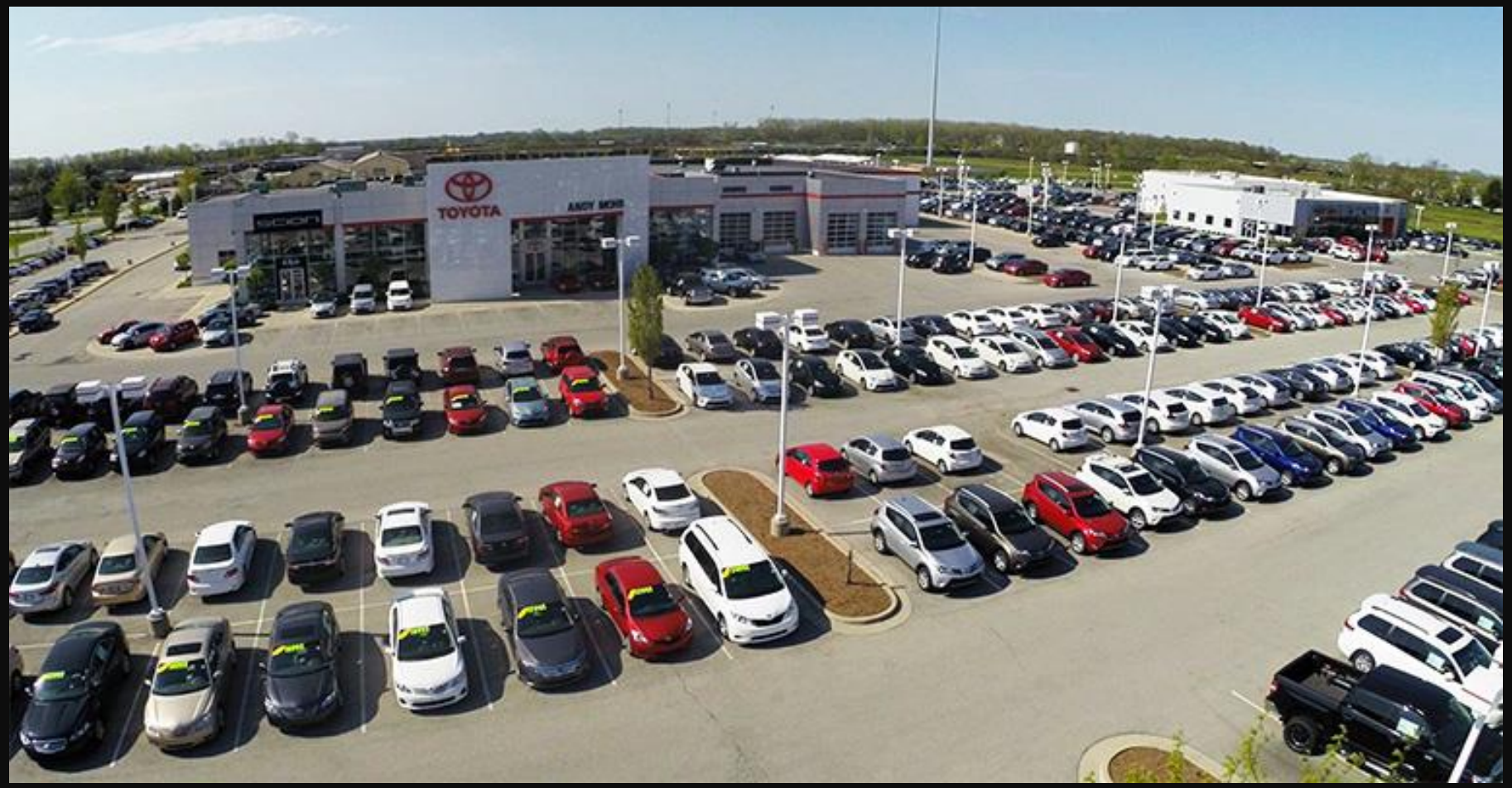
<u>Simple</u> linear regression example



Load the real estate data as <u>realestate</u>

- ✔ Model house price as DV and lotsize as IV
- ✔ Predict the price of houses with lot sizes 500, 1k, 2k square feet
- ✔ How much variance of house prices does this model explain?
- ✔ Find the "top" 5 outliers, remove them from the data and re-run the model. Did the variance explained improve? Why?
- ✔ Now use a log of the DV and IV by using the following:
  lm(log(price) ~ log(lotsize), data=realestate)
  Did the variance explained improved? Why?

<u>Multiple</u> linear regression example



Going back to the realestate dataframe

- ✔ Find the best multiple linear model to predict house prices.
- ✔ Your answer should note the IV used and why you have used them, test plots, removing outliers (if needed), transformations (if needed), and examining multicollinearity.